Solution sheet

Sequence Alignments and Phylogeny
Bioinformatics Leipzig

WS 13/14

1 Biological Background

1.1 Which of the following are pyrimidines?

Cytosine and Thymine are pyrimidines (number 2)

1.2 Which of the following contain phosphorus atoms?

DNA and RNA contain phosphorus atoms (number 2).

1.3 Which of the following contain sulfur atoms?

Methionine contains sulfur atoms (number 3).

1.4 Which of the following is not a valid amino acid sequence?

There is no amino acid with the one letter code 'O', such that there is no valid amino acid sequence 'WATSON' (number 4).

1.5 Which of the following 'one-letter' amino acid sequence corresponds to the sequence Tyr-Phe-Lys-Thr-Glu-Gly?

The amino acid sequence corresponds to the one letter code sequence YFKTEG (number 1).

1.6 Consider the following DNA oligomers. Which to are complementary to one another? All are written in the 5' to 3' direction (i.TTAGGC ii.CGGATT iii.AATCCG iv.CCGAAT)

CGGATT (ii) and AATCCG (iii) are complementary (number 2).

2 Pairwise Alignments

2.1 Needleman-Wunsch Algorithm

Given the alphabet $B = \{A, C, G, T\}$, the sequences $s = ACGCA$ and $p = ACCG$ and the following scoring matrix D:

	A	C	T	G	-
A	3	-1	-1	-1	-2
C	-1	3	-1	-1	-2
T	-1	-1	3	-1	-2
G	-1	-1	-1	3	-2
-	-2	-2	-2	-2	0

1. What kind of scoring function is given by the matrix D, similarity or distance score?
2. Use the Needleman-Wunsch algorithm to compute the pairwise alignment of s and p.
3. Do the backtracking step and write down the resulting optimal global alignment of the two sequences.

Solution:

The scoring matrix defines a similarity measure since the matches have a high positive score, thus, the overall score is a maximization function. The matrix shows the DP matrix of the Needleman-Wunsch algorithm:

p \ s	_	A	C	G	C	A
_	**0**	-1	-2	-3	-4	-5
A	-1	**3**	2	1	0	3
C	-2	2	**6**	5	8	7
C	-3	1	**9**	8	11	10
G	-4	0	8	**12**	**11**	**10**

The bold numbers show the path from the backtracking step. In this way we get the following pairwise alignment:

s	A	C	_	G	C	A
p	A	C	C	G	_	_

3 MSA and Profile Alignments

3.1 Given the following multiple sequence alignment (without gaps!)

col.no/seq.no	1	2	3	4	5	6	7	8	9	10	11	12
1	A	A	C	T	C	G	A	C	T	T	C	G
2	A	T	G	A	T	G	A	C	T	T	G	G
3	A	T	G	A	T	C	A	G	T	G	C	C
4	G	T	C	A	T	C	A	T	G	G	G	C
5	G	A	C	A	T	C	A	C	C	T	C	C

1. Construct the profile P for the MSA.

2. Construct the consensus sequence K.

3. When aligning as many sequences, why is working with profiles more suitable than using the consensus sequence?

4. Given the sequence S = GAGACGACTGGG. Calculate the score for the alignment of K and S (using the edit distance) and the score for the alignment of S and P (using the score below and the edit distance).

   Edit distance (for columns i,j):

   $$d(x_i, y_j) = \begin{cases} 1 & \text{if } x_i \neq y_j \\ 0 & \text{if } x_i = y_j \end{cases}$$

   Profile scoring (for columns i,j):

   $$B = \{A, T, C, G\}$$

   $$sc(s_i, p_j) = \sum_{b \in B} d(b, s_i) * p_{b,j}$$

   Hints:
   To get the overall score for the whole sequence compute the sum over the scores for each column i in the alignment.
   Since there are no gaps in the alignment, the column indices i and j are the same (i=j).

Solution:

1. Profile P (without gaps):

column \ base	A	C	G	T
1	3/5	0	2/5	0
2	2/5	0	0	3/5
3	0	3/5	2/5	0
4	4/5	0	0	1/5
5	0	1/5	0	4/5
6	0	3/5	2/5	0
7	1	0	0	0
8	0	3/5	1/5	1/5
9	0	1/5	1/5	3/5
10	0	0	2/5	3/5
11	0	3/5	2/5	0
12	0	3/5	2/5	0

2. Consensus sequence K: ATCATCACTTCC

3. With profiles an (almost) uniform distribution of the bases can be shown and used better than with a consensus sequence, which just takes one out of the bases when they are equally distributed.

4. Scores
   Aligning S with K:

S	G	A	G	A	C	G	A	C	T	G	G	G
K	A	T	C	A	T	C	A	C	T	T	C	C
Score	1	1	1	0	1	1	0	0	0	1	1	1

   The overall score for the alignment of S and K, $Sc(S, K) = 8$.

   Aligning S with P:

Column	1	2	3	4	5	6	7	8	9	10	11	12
S	G	A	G	A	C	G	A	C	T	G	G	G
Score	3/5	3/5	3/5	1/5	4/5	3/5	0	2/5	2/5	3/5	3/5	3/5

   The overall score for the alignment of S and P, $Sc(S, P) = 6$.

   Example for the sum for column 1:

$$\begin{aligned} Sc(s_1, p_1) &= d(A, G) * p_{A,1} + d(C, G) * p_{C,1} + d(G, G) * p_{G,1} + d(T, G) * p_{T,1} \\ &= 1 * \frac{3}{5} + 1 * 0 + 0 * \frac{2}{5} + 1 * 0 = \frac{3}{5} \end{aligned}$$

5. Conclusion: The score for the alignment of S with P is smaller which is in this case the better score. Thus, it can be seen that the profile alignment gives a better score because it takes into account not only the bases which appear the most. To get more suitable results, the edit distance can be replaced by another distance measure, which uses also values other than just 0 or 1.

4 Phylogeny

1. What does UPGMA do? (input/output)
2. What is the difference between UPGMA and WPGMA? (formula)

1. Given a matrix which shows distances or similarities between species, a phylogenetic tree can be constructed. In the phylogenetic tree, the leaves are the species, and either the inner nodes or the edges are labeled with the distances.
2. UPGMA is the unweighted case, where the distances between the cluster are *equally weighted (=unweighted)*. The WPGMA, thus weighted case, calculates the distances by weighting the clusters depending on their size.
   Let $X = A \cup B$ and $K = \{K_1, ..., K_n\}$ the set of all clusters except X, then

$$d_{X,K_i} = \frac{d_{AK_i} + d_{BK_i}}{2} \textit{ (WPGMA)}$$

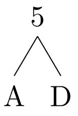
$$d_{X,K_i} = \frac{|A| * d_{AK_i} + |B| * d_{BK_i}}{|A| + |B|} \textit{ (UPGMA)}$$

4.1 Construct the a) UPGMA tree and b) the WPGMA tree for the following distance matrix

	A	B	C	D
A	0	9	7	5
B	9	0	8	10
C	7	8	0	8
D	5	10	8	0

a) UPGMA algorithm:

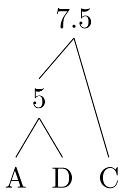
1. minimal $D_{ij} (i \neq j) : D_{AD} = 5$.
   Put A and D together in the tree.



Update distance matrix:

	AD	B	C
AD	0	9,5	7,5
B	9,5	0	8
C	7,5	8	0

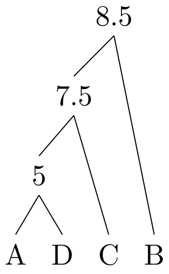
2. minimal $D_{ij}(i \neq j) : D_{ADC} = 7,5.$
   Put C next to AD in the tree.



Update distance matrix:

	ADC	B
ADC	0	8,5
B	8,5	0

3. Since B is the last node to add, put B next to ADC in the tree.

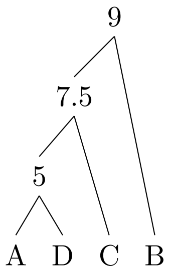


b) WPGMA algorithm:
Since in the first steps, the clusters have size 1, there is no difference between the algorithms. Just in the last step, there is a difference:

Update distance matrix, thus, compute the distance between cluster ADC and B:

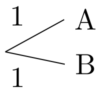
$$d_{ADC,B} = \frac{2 * 9.5 + 1 * 8}{2 + 1} = 9$$

	ADC	B
ADC	0	9
B	9	0



4.2 Compute the WPGMA tree for the following distance matrix. Write the resulting distances at the edges of the tree.

1. Group together these OTUs (operational taxonomical units) for which the distance is minimal. In this case group together A and B. The depth of the divergence can be computed by the WPGMA formula.



2. Compute the distance from cluster (A, B) to each other OTU and update the distance matrix.

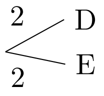
$$\begin{aligned} d_{(AB)C} &= (d_{AC} + d_{BC})/2 = 4 \\ d_{(AB)D} &= (d_{AD} + d_{BD})/2 = 6 \\ d_{(AB)E} &= (d_{AE} + d_{BE})/2 = 6 \\ d_{(AB)F} &= (d_{AF} + d_{BF})/2 = 8 \end{aligned}$$

	(AB)	C	D	E	F
(AB)	0				
C	4	0			
D	6	6	0		
E	6	6	4	0	
F	8	8	8	8	0

Repeat steps 1 and 2 until all OTUs are clustered (repeat until $N = 2$)

$N := N - 1 = 5$

3. Group together these OTUs for which the distance is minimal, e.g. group together D and E. Alternatively, (AB) could be grouped with C.



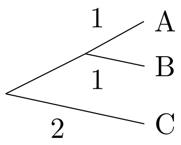
4. Compute the distance from cluster (DE) to each other OTU (cluster)

$$\begin{aligned} d_{(DE)(AB)} &= (d_{D(AB)} + d_{E(AB)})/2 = 6 \\ d_{(DE)C} &= (d_{DC} + d_{EC})/2 = 6 \\ d_{(DE)F} &= (d_{DF} + d_{EF})/2 = 8 \end{aligned}$$

	(AB)	C	(DE)	F
(AB)	0			
C	4	0		
(DE)	6	6	0	
F	8	8	8	0

$N := N - 1 = 4$

5. Group together these OTUs for which the distance is minimal, e.g. group (AB) and C



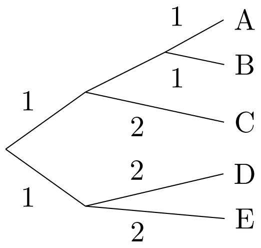
6. Compute the distance from cluster (ABC) to each other OTU (cluster)

$$\begin{aligned} d_{(ABC)(DE)} &= (d_{(AB)(DE)} + d_{C(DE)})/2 = 6 \\ d_{(ABC)F} &= (d_{(AB)F} + d_{CF})/2 = 8 \end{aligned}$$

	(ABC)	(DE)	F
(ABC)	0		
(DE)	6	0	
F	8	8	0

$N := N - 1 = 3$

7. Group together these OTUs for which the distance is minimal, e.g. group (ABC) and (DE)

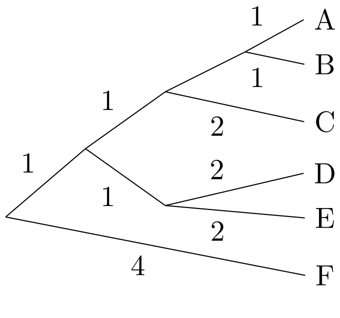


8. Compute the distance from cluster (A, B, C, D, E) to OTU F

$$d_{(ABCDE)F} = (d_{(ABC)F} + d_{(DE)F})/2 = 8$$

	((ABC)(DE))	F
((ABC)(DE))	0	
F	8	0

$N := N - 1 = 2$
The last two clusters can easily be put next to each other in the tree.


1
A
B
1
1
2
C
1
2
D
1
E
2
4
F